УДК 519.237.8+510.22

Д.А. Вятченин, А.В. Доморацкий
Объединенный институт проблем информатики НАН Беларуси, г. Минск, Беларусь
viattchenin@mail.ru
УП «Геоинформационные системы» НАН Беларуси, г. Минск, Беларусь
a.damaratski@gmail.com

Построение нечеткого c-разбиения в случае неустойчивой кластерной структуры множества объектов

Предложен метод кластеризации объектов с варьирующимися в интервале значениями признаков в случаях неустойчивой кластерной структуры множества объектов. Приводятся результаты вычислительного эксперимента.

Введение

В задачах автоматической классификации динамических объектов, то есть объектов, признаки которых могут изменять свои значения с течением времени или при наличии внешних воздействий [1], именуемых также задачами динамической кластеризации, признаки $\hat{x}^{t_1}$, $t_1 = 1,\ldots,m_1$ объектов $x_i \in X$ могут принимать значения в непрерывном интервале безотносительно к моменту измерения соответствующей характеристики объекта, так что каждый признак $\hat{x}^{t_1}$, $t_1 = 1,\ldots,m_1$ для объекта x_i, $i = 1,\ldots,n$ представляет собой интервал значений $[\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)}]$. Кластерная структура исследуемой совокупности, состоящей из подобных объектов, также является динамической, и зависит от значений признаков в момент классификации. На содержательном уровне задача построения устойчивой кластерной структуры в [1] формулируется следующим образом: найти такое априори неизвестное число c областей признакового пространства $\Re^m$, в которых отображаются кластеры при различных значениях принимаемых объектами исследуемой совокупности X признаков $\hat{x}^{t_1}$, $t_1 = 1,\ldots,m_1$, варьирующихся в интервале $[\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)}]$. В свою очередь, перед решением указанной задачи в первую очередь необходимо установить тип динамических изменений кластерной структуры, для чего в [1] определены понятия устойчивой, квазиустойчивой и неустойчивой кластерной структуры. Если при изменении в соответствующем интервале $[\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)}]$ значений признаков $\hat{x}^{t_1}$, $t_1 = 1,\ldots,m_1$ объектов $x_i \in X$ исследуемой совокупности число c кластеров $\{A^1,\ldots,A^c\}$ не изменяется и не изменяются координаты их прототипов $\{\bar{\tau}^1,\ldots,\bar{\tau}^c\}$, то структура, образуемая кластерами $\{A^1,\ldots,A^c\}$, называется *устойчивой*; если с изменением значений признаков объектов число c кластеров $\{A^1,\ldots,A^c\}$ не изменяется, но изменяются координаты их прототипов $\{\bar{\tau}^1,\ldots,\bar{\tau}^c\}$, то соответствующая кластерная структура именуется *квазиустойчивой*, а если при изменении значений признаков

наблюдаемых объектов $x_i \in X$ изменяется число c кластеров, то кластерная структура является *неустойчивой*. В [1] представлен метод определения типа кластерной структуры совокупности объектов с варьирующимися в интервале значениями признаков, в основе которого лежит D-AFC-TC-алгоритм [2] построения распределения объектов по априори неизвестному числу нечетких α-кластеров.

В случае, когда кластерная структура, образуемая объектами исследуемой совокупности, является неустойчивой, ей соответствуют такие типы динамических изменений, как образование новых кластеров, слияние кластеров, расщепление кластеров и элиминация кластеров, а в случае квазиустойчивости кластерной структуры число кластеров не изменяется, однако имеет место дрейф прототипов кластеров, и, как продемонстрировано в [1], изменяются типичные точки кластеров. В отличие от ситуации неустойчивой кластерной структуры, где ее изменения носят скачкообразный характер, в ситуации квазиустойчивой кластерной структуры изменения носят непрерывный и, как следствие, латентный характер. Метод построения распределения объектов по классам в случае квазиустойчивости кластерной структуры представлен в работе [3].

Целью предпринятого исследования является решение задачи построения нечеткого c-разбиения множества объектов, описываемых динамическими признаками, в случае, когда кластерная структура является неустойчивой. Основой предлагаемого метода является реляционный подход к решению задачи нечеткой кластеризации в сочетании с представлением объектов исследуемой совокупности как интервально-значных нечетких множеств и последующим построением матрицы коэффициентов различия на соответствующем универсуме.

Реляционные методы оптимизационного подхода к решению задачи нечеткой кластеризации

При решении задач динамической кластеризации могут использоваться подходы, основанные на методах нечеткой и возможностной кластеризации [2], в которых результатом классификации является не только отнесение i-го объекта исследуемой совокупности $X = \{x_1, \ldots, x_n\}$ к l-му классу A^l, $l = 1, \ldots, c$ но и указание функции принадлежности $\mu_{li} \in [0,1]$, $l = 1, \ldots, c$, $i = 1, \ldots, n$ с которой объект $x_i \in X$, $\forall i = 1, \ldots, n$ принадлежит тому или иному нечеткому кластеру $A^l, l = 1, \ldots, c$.

В нечетких методах автоматической классификации оптимизационного направления [4] нечеткая кластеризация понимается как разбиение классифицируемой совокупности объектов $X = \{x_1, ..., x_n\}$ на семейство его нечетких множеств, так что в качестве входного параметра в существующих оптимизационных нечетких методах нечеткой кластеризации, как правило, задается число нечетких кластеров c, причем при данном подходе под нечетким кластером может пониматься любое нечеткое множество, определенное на универсуме. Нечеткие множества A^l, $l = 1, \ldots, c$ с соответствующими функциями принадлежности $\mu_{1i}, \ldots, \mu_{ci}$ образуют нечеткое c-разбиение, если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^{c} \mu_{li} = 1$, и нечеткая модификация задачи кластеризации в экстремальной постановке заключается в нахождении экстремума некоторого функционала $Q(P)$ на множестве Π всех возможных нечетких c-разбиений P классифицируемого множества объектов X, что описывается формулой $Q(P) \to \underset{P \in \Pi}{extr}$.

Если исходные данные представлены матрицей вида «объект – свойство» $\hat{X}_{n\times m_1} = [\hat{x}_i^{t_1}]$, $i = 1,\ldots,n$, $t_1 = 1,\ldots,m_1$ задача классификации заключается в минимизации критерия, в который введена некоторая метрика, и типичным примером подобных функционалов является критерий, предложенный Дж. Данном [5] и позже обобщенный Дж. Беждеком [6]

$$Q_{FCM}(P,\overline{\mathrm{T}}) = \sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i - \overline{\tau}^l\right\|^2, \qquad (1)$$

процедура минимизации которого широко известна под обозначением FCM-алгоритма. Критерий (1), где символом $\overline{\tau}^l$ обозначен прототип l-го нечеткого кластера, а символом γ – задаваемый исследователем коэффициент нечеткости классификации, такой, что $1 < \gamma < \infty$, послужил основой для целого семейства функционалов и соответствующих им нечетких кластер-процедур, которые подробно рассматриваются в [4].

В случае представления данных в виде матрицы «объект – объект» $d_{n\times n} = [d(x_i, x_j)]$, $i, j = 1,\ldots,n$, элементами которой являются попарные коэффициенты различия, то для классификации объектов исследуемой совокупности используются так называемые реляционные процедуры, основанные на минимизации соответствующих функционалов, которые будут рассмотрены подробнее.

Одним из первых примеров такого рода функционалов может послужить критерий М. Рубенса [7]

$$Q_{FNM}(P) = \sum_{l=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_{li}^2\mu_{lj}^2 d(x_i, x_j), \qquad (2)$$

процедура минимизации которого именуется в литературе FNM-алгоритмом. В выражении (2) символом $d(x_i, x_j)$ обозначается коэффициент различия между объектами x_i и x_j, $i, j = 1,\ldots,n$ исследуемой совокупности X, $card(X) = n$, а μ_{li} – значение принадлежности i-го объекта l-му нечеткому кластеру.

В работе [8] Р.Дж. Гэтвэем, Дж. Беждеком и Дж.В. Дэвенпортом предложен реляционный аналог критерия (1), процедура минимизации которого получила обозначение RFCM-алгоритма. Метод основан на минимизации критерия вида

$$Q_{RFCM}(P) = \sum_{l=1}^{c}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\mu_{li}^{\gamma}\mu_{lj}^{\gamma}d(x_i, x_j)\right)\Big/\left(2\sum_{k=1}^{n}\mu_{lk}^{\gamma}\right)\right), \qquad (3)$$

с учетом того, что для всех $i, j = 1,\ldots,n$ имеет место выполнение условия евклидовой согласованности

$$d(x_i, x_j) = \left\|x_i - x_j\right\|^2, \qquad (4)$$

то есть матрица попарных расстояний $d_{n\times n} = [d(x_i, x_j)]$, $i, j = 1,\ldots,n$ описывает геометрическую структуру исследуемой совокупности в некотором m_1-мерном евклидовом пространстве $\Re^{m_1}$. Необходимо указать, что в работе [9] предложена модификация RFCM-алгоритма, не требующая выполнения условия (4), в силу чего получившая обозначение NERFCM-алгоритма.

В работе [10] М.П. Уиндхемом был предложен AP-алгоритм, минимизирующий критерий

$$Q_{AP}(P) = \sum_{l=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_{li}^2\nu_{lj}^2 d(x_i, x_j), \qquad (5)$$

при ограничениях на значения весов прототипов

$$\sum_{j=1}^{n}\nu_{lj} = 1,\ l = 1,\ldots,c,\ j = 1,\ldots,n, \qquad (6)$$

представляющий собой «наиболее сильное обобщение среди всех оптимизационных методов решения нечеткой модификации задачи кластерного анализа, и с содержательной точки зрения, минимизация (5) отыскивает оптимальное отклонение нечетких принадлежностей объектов от нечетких центров кластеров» [11].

Таким образом, нечеткой является не только степень принадлежности объекта класссу, но и сам представитель класса – каждый элемент классифицируемого множества $X = \{x_1, ..., x_n\}$ в различной степени может быть, с одной стороны, прототипом того или иного класса, а с другой – просто элементом этого и всех остальных классов. Результатом работы алгоритма, доставляющего минимум функционалу (5), являются матрица разбиения $P_{c\times n} = [\mu_{li}]$, где $l = 1,\ldots,c, i = 1,\ldots,n$, и матрица весов прототипов $K_{c\times n} = [\nu_{li}]$, где также $l = 1,\ldots,c$, $i = 1,\ldots,n$, и совместный анализ которых позволяет детально исследовать структуру множества объектов X.

В свою очередь, является ARCA-алгоритм, предложенный в [12], состоит в нахождении экстремума функционала

$$Q_{ARCA}(P, \mathrm{T}) = \sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left(\sqrt{\sum_{j=1}^{n}\left(d(x_i, x_j) - d(x_j, \tau^l)\right)^2}\right)^2, \quad (7)$$

где для показателя нечеткости классификации γ, как и в критерии (1), выполняется условие $1 < \gamma < \infty$, $d(x_i, x_j)$ – коэффициент различия объектов x_i и x_j, а значение $d(x_j, \bar{\tau}^l)$ выражает различие между объектом x_j и прототипом $\bar{\tau}^l$ нечеткого кластера $A^l, l = 1,\ldots,c$, для некоторого объекта $x_i \in X$. Результатом работы кластер-процедуры будет матрица $P_{c\times n} = [\mu_{li}]$ нечеткого c-разбиения P^*, а также матрица $D_{c\times n} = [d_{li}]$, $d_{li} = d(x_i, \bar{\tau}^l)$ расстояний объектов $x_i, i = 1,\ldots,n$ от прототипов $\bar{\tau}^l, l = 1,\ldots,c$, смысл которых во многом схож со смыслом вспомогательной функции принадлежности ν_{lj} в критерии (5), однако, в отличие от AP-алгоритма, в предложенном в [12] методе на значения d_{li}, $l = 1,\ldots,c$, $i = 1,\ldots,n$ не налагается условия $\sum_{i=1}^{n} d_{li} = 1$; кроме того, на значения $d(x_i, x_j)$ в матрице исходных данных не налагается никаких ограничений.

В силу того, что c является параметром любой оптимизационной кластер-процедуры, одной из главных проблем при использовании оптимизационных методов является проблема обоснования числа c нечетких кластеров, встающая наиболее остро, когда исследователю число классов c вообще неизвестно, для решения которой были предложены различные показатели, характеризующие получаемое при использовании того или иного алгоритма нечеткое c-разбиение $P^* = \{A^1,\ldots,A^c\}$. В частности, для FCM-алгоритма и его модификаций различными исследователями был введен ряд показателей, наиболее известными из которых являются:

– коэффициент разбиения

$$V_{pc}(P) = \frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^2, \quad (8)$$

предложенный Дж. Данном [13] и для которого решение задачи определения оптимального числа классов в P^* отыскивается в виде $\max_c\left(V_{pc}(P)\right), c = 2,\ldots,n-1$;

– энтропия разбиения

$$V_{pe}(P) = -\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\left|\mu_{li}\cdot\ln\mu_{li}\right|, \quad (9)$$

предложенная Дж. Беждеком, М.П. Уиндхемом и Р. Эрлихом [14], так что оптимальному числу классов в P^* соответствует $\min_c\left(V_{pe}(P)\right), c = 2,\ldots,n-1$;

– индекс разделимости

$$V_{si}(P) = \frac{\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^2\left\|x_i - \overline{\tau}^l\right\|^2}{\min_{l\neq k}\left\|\overline{\tau}^l - \overline{\tau}^k\right\|^2}, \tag{10}$$

предложенный Х.Л. Хи и Ж. Бени [15], где оптимальное число классов в P^* определяется, исходя из условия $\min_c\left(V_{si}(P)\right), c = 2,\ldots,n-1$. Как указывается в [16], для реляционных кластер-процедур приемлемыми оказываются непрямые индексы, такие, как (8) или (9); с другой стороны, в работе [12] отмечается, что, в силу возможности генерирования прототипов, показатель (10) также применим для использования с ARCA-алгоритмом.

Таким образом, применимость реляционных методов нечеткой кластеризации к классификации данных, описываемых матрицей вида «объект – свойство», обусловливается выбором метрики для получения матрицы вида «объект – объект». Следовательно, возникает вопрос о построении матрицы попарных коэффициентов различия для объектов, признаки которых описываются интервалами значений.

Методы обработки интервально-значных данных

При обработке интервально-значных данных представляется целесообразным прибегнуть к аппарату так называемых интервально-значных нечетких множеств, являющихся частным случаем нечетких множеств типа 2, которые также используются в задачах динамической кластеризации [17].

Пусть $X = \{x_1,...,x_n\}$ – множество объектов, так что каждый объект x_i описывается m_1 числом признаков и может быть представлен в виде вектора $x_i = (x_i^1,\ldots,x_i^{t_1},\ldots,x_i^{m_1})$, а каждый признак, в свою очередь $\hat{x}^{t_1}$, $t_1 = 1,\ldots,m_1$ описывается m_2 значениями, так что $\hat{x}_i^{t_1} = (\hat{x}_i^{t_1(1)},\ldots,\hat{x}_i^{t_1(t_2)},\ldots,\hat{x}_i^{t_1(m_2)})$. Таким образом, тринаправленные данные могут быть представлены в виде полиматрицы $\hat{X}_{n\times m_1\times m_2} = [\hat{x}_i^{t_1(t_2)}]$, $i = 1,\ldots,n$, $t_1 = 1,\ldots,m_1$, $t_2 = 1,\ldots,m_2$, которая может быть обработана с помощью обобщенной нормализации

$$x_i^{t_1(t_2)} = \frac{\hat{x}_i^{t_1(t_2)}}{\max_{i,t_2}\hat{x}_i^{t_1(t_2)}}, \tag{11}$$

или обобщенной унитаризации

$$x_i^{t_1(t_2)} = \frac{\hat{x}_i^{t_1(t_2)} - \min_{i,t_2}\hat{x}_i^{t_1(t_2)}}{\max_{i,t_2}\hat{x}_i^{t_1(t_2)} - \min_{i,t_2}\hat{x}_i^{t_1(t_2)}}, \tag{12}$$

так что каждый объект x_i, $i = 1,\ldots,n$ множества $X = \{x_1,\ldots,x_n\}$ может рассматриваться как нечеткое множество типа 2 с функциями принадлежности $x_i^{t_1(t_2)} = \mu_{x_i}(x^{t_1(t_2)})$, $i = 1,\ldots,n$; $t_1 = 1,\ldots,m_1$, $t_2 = 1,\ldots,m_2$, $x^{t_1(t_2)} = \mu_{t_1}(x^{t_2}) \in [0,1]$, $t_1 = 1,\ldots,m_1$, $t_2 = 1,\ldots,m_2$. В случае тринаправленных данных каждый объект x_i, $i = 1,\ldots,n$ описывается $X_{(i)m_1\times m_2} = [x_i^{t_1(t_2)}]$, $t_1 = 1,\ldots,m_1$, $t_2 = 1,\ldots,m_2$.

В ситуации интервальной неопределенности исследователь обладает информацией, что действительное значение $\hat{x}_i^{t_1}$ некоторого признака $\hat{x}^{t_1}$, $t_1 \in \{1,\ldots,m_1\}$ для объекта x_i, $i \in \{1,\ldots,n\}$ принадлежит некоторому интервалу и, если $t_2 \in \{\min,\max\}$, то $\hat{x}_i^{t_1} \in [\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)}]$. Таким образом, интервально-значные данные представляют собой частный случай тринаправленных данных, когда $m_2 = 2$, что описывается выражением $\hat{x}_i^{t_1} = (\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)})$. Очевидно, что в случае $\hat{x}^{t_1(1)} = \ldots = \hat{x}^{t_1(t_2)} = \ldots = \hat{x}^{t_1(m_2)}$ для всех $t_1 = 1,\ldots,m_1$, $\forall i = 1,\ldots,n$, исходные данные представляются обычной матрицей «объект – свойство», $\hat{X}_{n \times m_1} = [\hat{x}_i^{t_1}]$.

Для интервально-значных нечетких множеств Х. Юу и Х. Юаном в [18] был определен ряд мер близости, одна из которых определяется выражением

$$s_I(x_i,x_j) = 1 - \frac{1}{\sqrt[\lambda]{m_1}} \sqrt[\lambda]{\sum_{t_1=1}^{m_1} \left| \frac{\mu_{x_i}(x^{t_1(\min)}) + \mu_{x_i}(x^{t_1(\max)})}{2} - \frac{\mu_{x_j}(x^{t_1(\min)}) + \mu_{x_j}(x^{t_1(\max)})}{2} \right|^{\lambda}} \qquad (13)$$

для всех значений $1 \le \lambda < \infty$, так что применение к полученной матрице коэффициентов сходства $S_{n \times n} = [s_I(x_i,x_j)]$, $i,j = 1,\ldots,n$ операции дополнения

$$\mu_I(x_i,x_j) = 1 - s_I(x_i,x_j),\ \forall x_i,x_j,\ i,j = 1,\ldots,n, \qquad (14)$$

дает в результате матрицу нечеткого отношения несходства $I_{n \times n} = [\mu_I(x_i,x_j)]$.

Кроме того, П. Гжегожевским в [19] было предложено следующее расстояние между интервально-значными нечеткими множествами, основанное на метрике Хаусдорфа:

$$e_I(x_i,x_j) = \sqrt{\frac{1}{m_1} \sum_{t_1=1}^{m_1} \max\left\{ \left(\mu_{x_i}(x^{t_1(\min)}) - \mu_{x_j}(x^{t_1(\min)})\right)^2, \left(\mu_{x_i}(x^{t_1(\max)}) - \mu_{x_j}(x^{t_1(\max)})\right)^2 \right\}}. \qquad (15)$$

В свою очередь, учитывая, что интервально-значные нечеткие множества представляют собой частный случай нечетких множеств типа 2, где, как указывалось выше, $t_2 \in \{\min,\max\}$, то для построения матрицы исходных данных оказывается возможным применение к нормализованным интервально-значным данным обобщений расстояний для нечетких множеств типа 2, предложенным в [17]. В частности, обобщение квадрата относительного евклидова расстояния для случая интервально-значных нечетких множеств примет вид

$$\varepsilon_I(x_i,x_j) = \frac{1}{m_1} \sum_{t_1=1}^{m_1} \left(\frac{1}{2^2} \sum_{t_2 \in \{\min,\max\}} \left(\mu_{x_i}(x^{t_1(t_2)}) - \mu_{x_j}(x^{t_1(t_2)})\right)^2 \right), \qquad (16)$$

применение которого к x_i, $i = 1,\ldots,n$ также позволяет построить матрицу нечеткого отношения несходства $I_{n \times n} = [\mu_I(x_i,x_j)]$. Необходимо указать, что построенное с помощью формулы (16) нечеткое отношение, в общем, не является отношением несходства, так как сохраняет только свойство симметричности, но не является транзитивным и не удовлетворяет условию даже слабой антирефлексивности, так что обобщение квадрата относительного евклидова расстояния (16) представляет собой не расстояние, а меру различия.

Метод построения оптимального нечеткого с-разбиения

Задача построения нечеткого c-разбиения P^* на множестве $X = \{x_1,...,x_n\}$ динамических объектов заключается, таким образом, в построении функции распределения возможностей числа классов в искомом нечетком c-разбиении и построении матрицы коэффициентов различия с последующей обработкой некоторой реляционной кластер-

процедурой, с одновременным вычислением некоторого показателя валидности, при различных значениях числа классов $\hat{c}_g \in D_{\hat{\alpha}}$, для чего оказывается необходимым построение множества $D_{\hat{\alpha}} = Supp(D_{(\hat{\alpha})})$ наиболее возможных значений числа классов в искомом нечетком c-разбиении P^*. Метод построения функции распределения возможностей $\pi(\hat{c}_g)$ числа классов в искомой кластерной структуре, описанный в [20] и являющийся обобщением предложенного в [21] метода, заключается в построении нечеткого множества $D = \underset{\hat{t}_2}{\mathrm{S}}(\mu_{\hat{V}^{(\hat{t}_2)}}(\hat{c}_g))$, представляющего собой нечеткое объединение нечетких множеств $\hat{V}^{(\hat{t}_2)} = \{\hat{c}_g, \mu_{\hat{V}^{(\hat{t}_2)}}(\hat{c}_g)\}$, $t_2 \in \{\min, \max\}$, построенных на основе результатов обработки матриц нормированных данных «объект – свойство» $X_{n\times m_1} = [x_i^{t_1(\min)}]$, $X_{n\times m_1} = [x_i^{t_1(\max)}]$ с помощью D-AFC-TC-алгоритма, с последующим построением нечеткого множества уровня $D_{(\hat{\alpha})} = \{(c_g \in D_\alpha, \mu_{D_{(\hat{\alpha})}}(\hat{c}_g) = \mu_D(\hat{c}_g))\}$, где $\hat{\alpha} = \min_{\hat{t}_2} \alpha^{(\hat{t}_2)}$, $t_2 \in \{\min, \max\}$ и $D_{\hat{\alpha}} = \{\hat{c}_g \in \hat{C} \mid \mu_D(\hat{c}_g) \geq \hat{\alpha}\}$, так что $D_{\hat{\alpha}} = Supp(D_{(\hat{\alpha})})$ и $\mu_{D_{(\hat{\alpha})}}(\hat{c}_g) = \pi(\hat{c}_g)$. Следует указать, что под нечетким объединением подразумевается некоторая S-норма, представляющая собой двухместную действительную функцию $\mathrm{S}:[0,1]\times[0,1] \to [0,1]$, удовлетворяющую условиям ограниченности, монотонности, коммутативности и ассоциативности. Если U – некоторый универсум, на котором определены нечеткие множества A и B, то наиболее распространенными S-нормами являются следующие [22]:

– операция максимума:

$$\mathrm{S}_1(\mu_A(u), \mu_B(u)) = \max(\mu_A(u), \mu_B(u)), \tag{17}$$

– алгебраическая сумма:

$$\mathrm{S}_2(\mu_A(u), \mu_B(u)) = \mu_A(u) + \mu_B(u) - \mu_A(u)\cdot\mu_B(u), \tag{18}$$

– ограниченная сумма:

$$\mathrm{S}_3(\mu_A(u), \mu_B(u)) = \min(1, \mu_A(u) + \mu_B(u)), \tag{19}$$

так что функция распределения возможностей, в конечном итоге, зависит от выбранного в качестве параметра D-AFC-TC-алгоритма расстояния между нечеткими множествами и выбранной S-нормы [20]. Таким образом, метод построения нечеткого c-разбиения множества $X = \{x_1, ..., x_n\}$, где значения признаков классифицируемых объектов описываются интервалами, может быть представлен в виде следующей последовательности шагов:

1. Производится нормировка исходных данных, представленных в виде матрицы интервально-значных данных $\hat{X}_{n\times m_1} = [\hat{x}_i^{t_1(\min)}, \hat{x}_i^{t_1(\max)}]$, $i = 1, \ldots, n$, $t_1 = 1, \ldots, m_1$, и полученная матрица нормированных данных разделяется на две матрицы, $X_{n\times m_1} = [x_i^{t_1(\min)}]$ и $X_{n\times m_1} = [x_i^{t_1(\max)}]$, каждая из которых обрабатывается D-AFC-TC-алгоритмом.

2. В соответствии с выбранной S-нормой строится функция распределения возможностей числа классов в искомом P^* и выделяется множество $D_{\hat{\alpha}} = Supp(D_{(\hat{\alpha})})$ наиболее возможных значений числа нечетких кластеров.

3. Путем применения к матрице нормализованных интервально-значных данных $X_{n\times m_1} = [x_i^{t_1(\min)}, x_i^{t_1(\max)}]$ выбранной метрики строится матрица нечеткого отношения несходства $I_{n\times n} = [\mu_I(x_i, x_j)]$, после чего производится ее обработка выбранного реляционного алгоритма нечеткой кластеризации и показателя валидности при всех значениях числа классов $\hat{c}_g \in D_{\hat{\alpha}}$.

Изложенный метод построения нечеткого c-разбиения представляет собой частный случай метода построения устойчивой кластерной структуры, предложенного в Работе [23].

Иллюстративный пример

Для иллюстрации предложенного метода были выбраны интервально-значные данные по восьми типам жиров – шести растительным и двум животным, рассмотренные М. Итино и Х. Ягути в [24], так что исследуемая совокупность должна расслаиваться на два класса. При проведении вычислительного эксперимента были выбраны данные по трем признакам [25], представленные в табл. 1, наиболее часто используемые в сравнительных исследованиях.

Таблица 1 – Интервально-значные данные М. Итино и Х. Ягути

Виды масел	Удельный вес	Содержание йода	Значение омыления
Льняное масло	0,930 – 0,935	170 – 204	118 – 196
Перилловое масло	0,930 – 0,935	192 – 208	188 – 197
Хлопковое масло	0,916 – 0,918	99 – 113	189 – 198
Кунжутное масло	0,920 – 0,926	104 – 116	187 – 193
Камелиевое масло	0,916 – 0,917	80 – 82	189 – 193
Оливковое масло	0,914 – 0,919	79 – 90	187 – 196
Говяжий жир	0,860 – 0,870	40 – 48	190 – 199
Свиной жир	0,858 – 0,864	53 – 77	190 – 202

Для применения предложенного метода была выбрана нормировка (12), а в качестве параметра D-AFC-TC-алгоритма – квадрат относительного евклидова расстояния. На рис. 1 а) – в) изображены функции распределения возможностей $\pi(\hat{c}_g)$ числа классов в P^*, построенные для S-норм (17) – (19) соответственно.

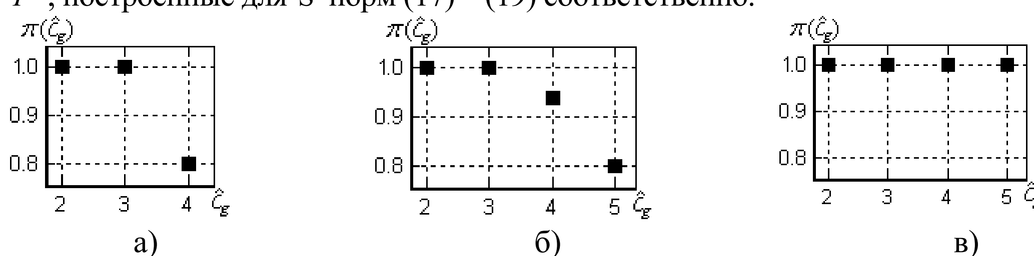


Рисунок 1 – Функции распределения возможностей числа классов

Вычислительный эксперимент проводился с использованием ARCA-алгоритма, а в качестве показателя валидности – коэффициента разбиения (8), значения которого на множестве $\hat{c}_g \in D_{\hat{\alpha}} = \{2,\ldots,4\}$ для всех рассмотренных мер различия (13) – (16) изображены на рис. 2 а) – в) соответственно.

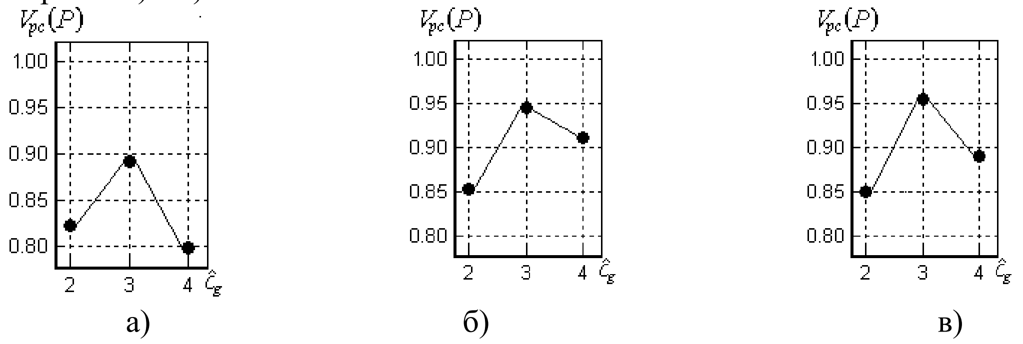


Рисунок 2 – Значения коэффициента разбиения при различных мерах несходства

Результаты кластеризации в виде матрицы нечеткого c-разбиения и матрицы расстояний от объектов до прототипов кластеров, полученные при использовании обобщения квадрата относительного евклидова расстояния (16) для построения матрицы попарных коэффициентов различия, представлены в табл. 2.

Таблица 2 – Результаты кластеризации данных М. Итино и Х. Ягути

Виды масел	Значения принадлежности			Расстояния до прототипов		
	Класс 1	Класс 2	Класс 3	Класс 1	Класс 2	Класс 3
Льняное масло	0,9542	0,0301	0,0157	0,1454	0,2462	0,6282
Перилловое масло	0,9270	0,0558	0,0172	0,0736	0,1466	0,5286
Хлопковое масло	0,0021	0,9962	0,0017	0,1778	0,0069	0,2017
Кунжутное масло	0,0207	0,9668	0,0125	0,1553	0,0091	0,2461
Камелиевое масло	0,0073	0,9845	0,0082	0,2321	0,0066	0,1751
Оливковое масло	0,0049	0,9896	0,0055	0,2251	0,0070	0,1786
Говяжий жир	0,0006	0,0013	0,9981	0,5965	0,1971	0,0090
Свиной жир	0,0006	0,0013	0,9981	0,5633	0,2035	0,0104

Результаты кластеризации, полученные с помощью расстояний (13) – (15) аналогичны приведенным в табл. 2. Таким образом, исследуемая совокупность разбивается на 3 класса – класс растительных жиров расслаивается на 2 субкластера, что соответствует приведенному в [24] результату иерархической классификации исследуемой совокупности. Следует также отметить, что выбор в качестве кластер-процедуры ARCA-алгоритма, минимизирующего критерий (7), обусловлен его возможностью обрабатывать матрицы коэффициентов различия, не удовлетворяющие условию антирефлексивности, что позволяет использовать для предобработки данных меру различия (16), а получаемый результат в виде матриц нечеткого c-разбиения $P_{c\times n} = [\mu_{li}]$ и расстояний от объектов до прототипов нечетких кластеров $D_{c\times n} = [d_{li}]$ позволяет углубить анализ.

С другой стороны, следует отметить, что применение аналогичного метода с использованием в качестве кластер-процедуры возможностного D-AFC(c)-алгоритма с соответствующим показателем валидности [23] строит распределение по 2 нечетким кластерам для каждой из рассмотренных мер (13), (15) и (16), однако удовлетворительный результат был получен только для меры близости (13).

Заключение

Результаты вычислительного эксперимента наглядно демонстрируют высокую эффективность предложенного метода построения нечеткого c-разбиения для оптимального, в смысле выбираемого показателя валидности, числа классов. Метод является гибким и эффективным с точки зрения возможности выбора кластер-процедуры, показателя валидности и меры различия между объектами; кроме того, помимо точности результатов классификации, достоинством предложенного метода является возможность обработки интервально-значных данных в полностью автоматическом режиме.

Литература

1. Вятченин Д.А. Анализ устойчивости кластерной структуры в задачах нестационарной кластеризации / Д.А. Вятченин // Доклады БГУИР. – 2009. – № 6. – С. 91-98.
2. Вятченин Д.А. Прямые алгоритмы нечеткой кластеризации, основанные на операции транзитивного замыкания и их применение к обнаружению аномальных наблюдений / Д.А. Вятченин // Искусственный интеллект. – 2007. – № 3. – С. 205-216.
3. Вятченин Д.А. Построение распределения по нечетким кластерам в случае квазиустойчивой кластерной структуры множества объектов / Д.А. Вятченин, А.В. Доморацкий // Доклады БГУИР. – 2010. – № 1. – С. 46-52.

4. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition / F. Höppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester : Wiley Intersciences, 1999. – 289 p.
5. Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters / J.C. Dunn // Journal of Cybernetics. – 1974. – Vol. 3. – P. 32-57.
6. Bezdek J.C. A convergence theorem for the fuzzy ISODATA clustering algorithms / J.C. Bezdek // IEEE Transactions on Pattern Analysis and Machines Intelligence. – 1980. – Vol. 2. – P. 1-8.
7. Roubens M. Pattern classification problems and fuzzy sets / M. Roubens // Fuzzy Sets and Systems. – 1978. – Vol. 1. – P. 239-253.
8. Hathaway R.J. On relational data versions of C-means algorithms / R.J. Hathaway, J.C. Bezdek, J.W. Dawenport // Pattern Recognition Letters. – 1996. – Vol. 17. – P. 607-612.
9. Hathaway R.J., Bezdek J.C. NERF C-means: non-Euclidean relational fuzzy clustering // Pattern Recognition. – 1994. – Vol. 27. – P. 429-437.
10. Windham M.P. Numerical classification of proximity data with assignment measures / M.P. Windham // Journal of Classification. – 1985. – Vol. 2. – P. 157-172.
11. Вятченин Д.А. Нечеткие методы автоматической классификации / Вятченин Д.А. – Минск : УП Технопринт, 2004. – 219 с.
12. Corsini P. A new fuzzy relational clustering algorithm based on fuzzy C-means algorithm / P. Corsini, B. Lazzerini, F. Marcelloni // Soft Computing. – 2005. – Vol. 9. – P. 439-447.
13. Dunn J.C. Well-separated clusters and the optimal fuzzy partitions / J.C. Dunn // Journal of Cybernetics. – 1974. – Vol. 4. – P. 95-104.
14. Bezdek J.C. Statistical parameters of cluster validity functionals / J.C. Bezdek, M.P. Windham, R. Ehrlich // International Journal of Computer Information Sciences. – 1980. – Vol. 9. – P. 323-336.
15. Xie X.L. A validity measure for fuzzy clustering / X.L. Xie, G. Beni // IEEE Transactions on Pattern Analysis and Machines Intelligence. – 1991. – Vol. 13. – P. 841-847.
16. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J.M. Keller, R. Krishnapuram, N.R. Pal. – New York: Springer Science, 2005. – 776 p.
17. Viattchenin D.A. An outline for a heuristic approach to possibilistic clustering of the three-way data / D.A. Viattchenin // Journal of Uncertain Systems. – 2009. – Vol. 3. – P. 64-80.
18. Ju H. Similarity measures on interval-valued fuzzy sets and application to pattern recognition / H. Ju, X. Yan // Fuzzy Information and Engineering / Ed. by D.Y. Cao. – Berlin : Springer-Verlag, 2007. – P. 875-883.
19. Grzegorzewski P. Distances between intuitionistic fuzzy sets and/or on interval-valued fuzzy sets based on Hausdorff metric / P. Grzegorzewski // Fuzzy Sets and Systems. – 2004. – Vol. 148. – P. 319-328.
20. Viattchenin D.A. An approach to constructing a possibility distribution for the number of fuzzy clusters / D.A. Viattchenin // Pattern Recognition and Information Processing: Proceedings of the 11th International Conference PRIP'2011(Minsk, Belarus, May 18 – 20, 2011) / Ed. by R. Sadykhov [et. al.] – Minsk : BSUIR, 2011. – P. 188-194.
21. Вятченин Д.А. Применение нечетких чисел для обоснования кластеров в методах нечеткой кластеризации / Д.А. Вятченин // Искусственный интеллект. – 2008. – № 3. – С. 523-533.
22. Нечеткие множества в моделях управления и искусственного интеллекта / А.Н. Аверкин, И.З. Батыршин, А.Ф. Блишун [и др.]; под ред. Д.А. Поспелова. – М. : Наука, 1986. – 312 с.
23. Viattchenin D.A. Constructing stable clustering structure for uncertain data set / D.A. Viattchenin // Acta Electrotechnica et Informatica. – 2011. – Vol. 11, № 3. (to appear)
24. Ichino M., Yaguchi H. Generalized Minkowski metrics for mixed feature-type data analysis / M. Ichino, H. Yaguchi // IEEE Transactions on Systems, Man, and Cybernetics. – 1994. – Vol. 25. – P. 698-708.
25. Sato-Ilic M. Innovations in Fuzzy Clustering: Theory and Applications / M. Sato-Ilic, L.C. Jain. – Heidelberg : Springer-Verlag, 2006. – 152 p.

Literatura

1. Vjatchenin D.A. Doklady BGUIR. №6. 2009. S. 91-98.
2. Vjatchenin D.A Iskusstvennyj intellekt. № 3. 2007. S. 205-216.
3. Vjatchenin D.A. Doklady BGUIR. №1 2010. S. 46-52.
4. Höppner F. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester: Wiley Intersciences. 1999. 289 p.
5. Dunn J.C. Journal of Cybernetics. Vol. 3. 1974. P. 32-57.
6. Bezdek J.C. IEEE Transactions on Pattern Analysis and Machines Intelligence. Vol. 2. 1980. P. 1-8.

7. Roubens M. Fuzzy Sets and Systems. Vol. 1. 1978. P. 239-253.
8. Hathaway R.J. Pattern Recognition Letters. Vol.17. 1996. P. 607-612.
9. Hathaway R.J. Pattern Recognition. Vol. 27. 1994. P. 429-437
10. Windham M.P. Journal of Classification. Vol 2. 1985. P. 157-172.
11. Vjatchenin D.A. Nechetkie metody avtomaticheskoj klassifikacii. Minsk.: UP Tehnoprint. 2004. 219 s.
12. Corsini P. Soft Computing. Vol 9. 2005. P. 439-447
13. Dunn J.C. Journal of Cybernetics. Vol 4. 1974. P. 95-104.
14. Bezdek J.C. International Journal of Computer Information Sciences. Vol. 9. 1980. P. 323-336.
15. Xie X.L. Transactions on Pattern Analysis and Machines Intelligence. Vol. 13. 1991. P. 841-847.
16. Bezdek J.C. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. New York: Springer Science. 2005. 776 p.
17. Viattchenin D.A. Journal of Uncertain Systems. Vol. 3. 2009. P. 64-80.
18. Ju H. Fuzzy Information and Engineering. Berlin: Springer-Verlag. 2007. P. 875-883.
19. Grzegorzewski P. Fuzzy Sets and Systems. Vol 148. 2004. P. 319-328.
20. Viattchenin D.A. Pattern Recognition and Information Processing: Proceedings of the 11th International Conference PRIP’2011(Minsk, Belarus, May 18-20, 2011). Minsk: BSUIR. 2011. P. 188-194.
21. Vjatchenin D.A. Iskusstvennyj intellekt. № 3. 2008. S. 523-533.
22. Averkin A.N. Nechetkie mnozhestva v modeljah upravlenija i iskusstvennogo intellekta. M.: Nauka. 1986. 312 s.
23. Viattchenin D.A. Acta Electrotechnica et Informatica. Vol. 11. №.3. 2011 (to appear)
24. Ichino M. IEEE Transactions on Systems, Man, and Cybernetics. Vol 25.1994. P. 698-708.
25. Sato-Ilic M. Innovations in Fuzzy Clustering: Theory and Applications. Heidelberg: Springer-Verlag. 2006. 152 p.

Д.А. Вятченін, А.В. Доморацький

Побудова нечіткого с-розбиття у разі нестійкої кластерної структури безлічі об’єктів

Запропонований метод кластеризації об’єктів із значеннями ознак, що варіюються в інтервалі, у випадках нестійкої кластерної структури безлічі об’єктів. Наводяться результати обчислювального експерименту.

D.A. Viattchenin, A.V. Damaratski

Constructing Fuzzy C-Partition in Case of Unstable Cluster Structure of Set of Objects

A method of clustering of objects for varying in an interval attributes values in a case of the unstable cluster structure of the set of objects is proposed. Results of the numerical experiment are presented.

Статья поступила в редакцию 22.06.2011.